Can a Proactive AI Tutor Improve Learning in Real Classroom Math Practice?

Experimental Evidence from NUMI

Philip Oreopoulos (University of Toronto)

Michael Liut (University of Toronto)

Alp Sungu (Wharton School)

Nina Low (NBER and Yale University)

Motivation

- AI has the potential to mimic the 1 student 1 teacher model (like tutoring but less costly, like CAL, but more interactive)
- Challenge is not tech but behavioral
- Most students don't feel motivated or incentivized to use and AI tutor set up as a chatbot in the corner of the screen.
- In practice, students often ignore them, ask for the answer, or engage very little.
- NUMI is built around the idea that the main problem is not only AI quality.
- The main problem is how to design the environment and guardrails so students use AI in educationally productive ways.

Contribution

1. A clean test of the **added value of LLM technology** relative to a standard computer assisted learning platform with worked solutions.
2. A tutoring design built explicitly around **engagement frictions**: students who do not like typing, do not want to talk to a bot, or just want to move on.
3. A research platform that can be used for future **A/B testing of AI features** as the technology evolves.

Roadmap

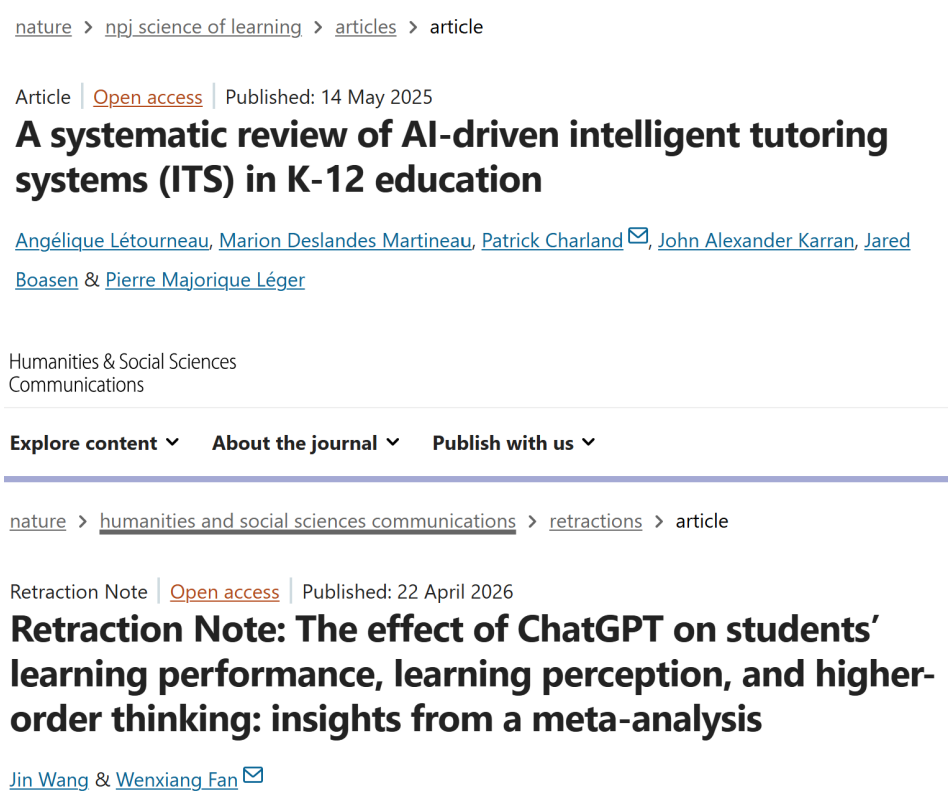
1. Motivation and past evidence
2. What NUMI is designed to do differently
3. Study design and sample
4. How students actually used NUMI
5. Main results: pace, correctness, and post-mistake outcomes
6. Delayed test evidence
7. Interpretation and implications

Previous evidence is mixed

- **Tutoring** remains the gold standard, but is costly and challenging to scale.
- **Tutor CoPilot** suggests AI can improve the quality of human tutoring by assisting tutors in real time.
- **ChatGPT-generated advice** can perform about as well as human-authored advice or better in online math settings.
- **TeachGPT** reports large gains in a college physics setting, but against lecture-based instruction rather than against an existing digital practice platform.
- **Bastani et al.** show that unstructured access to GPT-4 can harm learning in high school math; carefully prompted AI mitigates the harm, but still does not improve performance over the control.
- **LearnLM / Eedi** finds promising post-mistake and next-topic results in supervised online tutoring, but in a tutoring environment rather than inside a classroom mastery-practice workflow.

Selected references: Wang et al. (2024) Tutor CoPilot; Pardos and Bhandari (2024); Kestin et al. (2025); Bastani et al. (2025); LearnLM Team and Eedi (2025).

"meta-analyses" shaky



nature › npj science of learning › articles › article

Article | Open access | Published: 14 May 2025

A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education

Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen & Pierre Majorique Léger

Humanities & Social Sciences Communications

Explore content | About the journal | Publish with us

nature › humanities and social sciences communications › retractions › article

Retraction Note | Open access | Published: 22 April 2026

Retraction Note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis

Jin Wang & Wenxiang Fan

"It's been more than 3.5 years since ChatGPT was commercially deployed, yet here we are today with virtually no empirical evidence that these tools can improve student learning outcomes in any enduring fashion. Remarkable."

How I read the past evidence

- It is surprisingly hard to generate learning gains from AI in real settings.
- One reason may be that students often do not engage with the tutor much.
- Another is that students are not well motivated or incentivised to use AI to learn
- The negative or null results are not just about prompt quality.
- They may also reflect the fact that the tutor is still a passive chatbot that students can ignore, misuse, or use minimally.

NUMI is built to test whether a more structured and more proactive AI tutor can produce value added, even if the gains are modest at first.

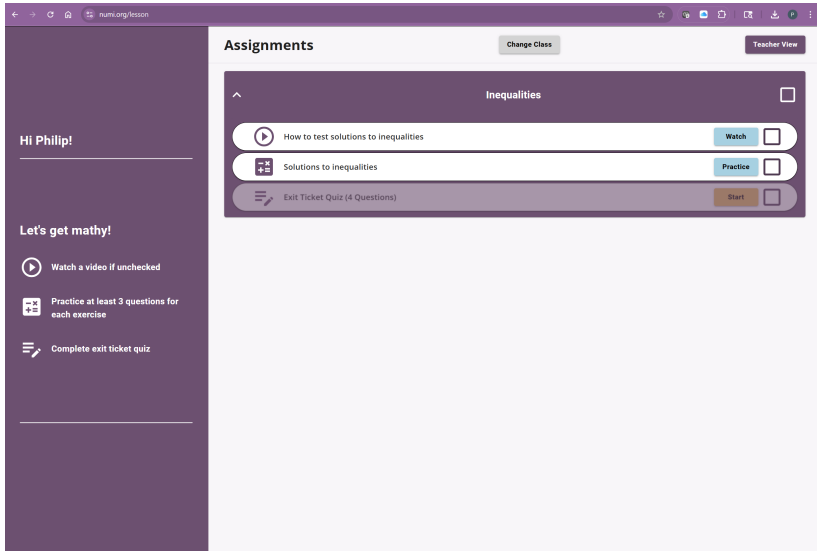
Key choice-architecture of NUMI

1. **Mastery structure**: students must follow a video/practice sequence before moving on.
2. **Button-driven interaction**: A/B, yes/no, and short-answer prompts make it easier to respond.
3. **Proactive support after mistakes**: students required to interact with NUMI after making a mistake.
4. **First-step hints**: Occasional "Help me get started" provides a low-friction way to nudge student forward or re-engage
5. **Voice**: NUMI reads its responses aloud.
6. Prompted towards reluctant student: non-patronizing, humorous, patient, focused on weak foundations and misconceptions

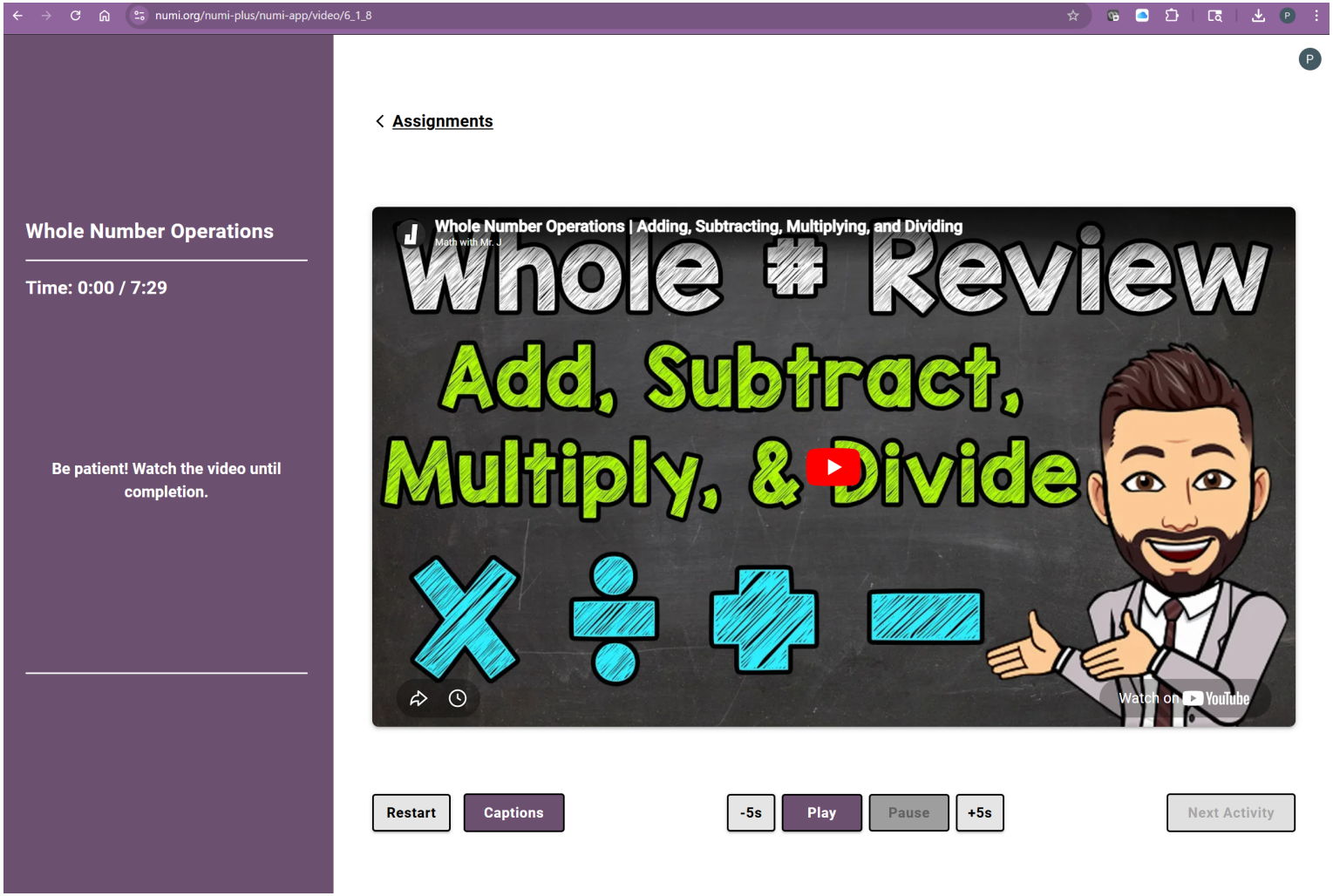
The assignment is intentionally simple

- Watch the video.
- Practice until getting three correct in a row.
- Complete the exit ticket.

NUMI is part of the lesson flow. It is not just a stand-alone chatbot floating beside the problem.

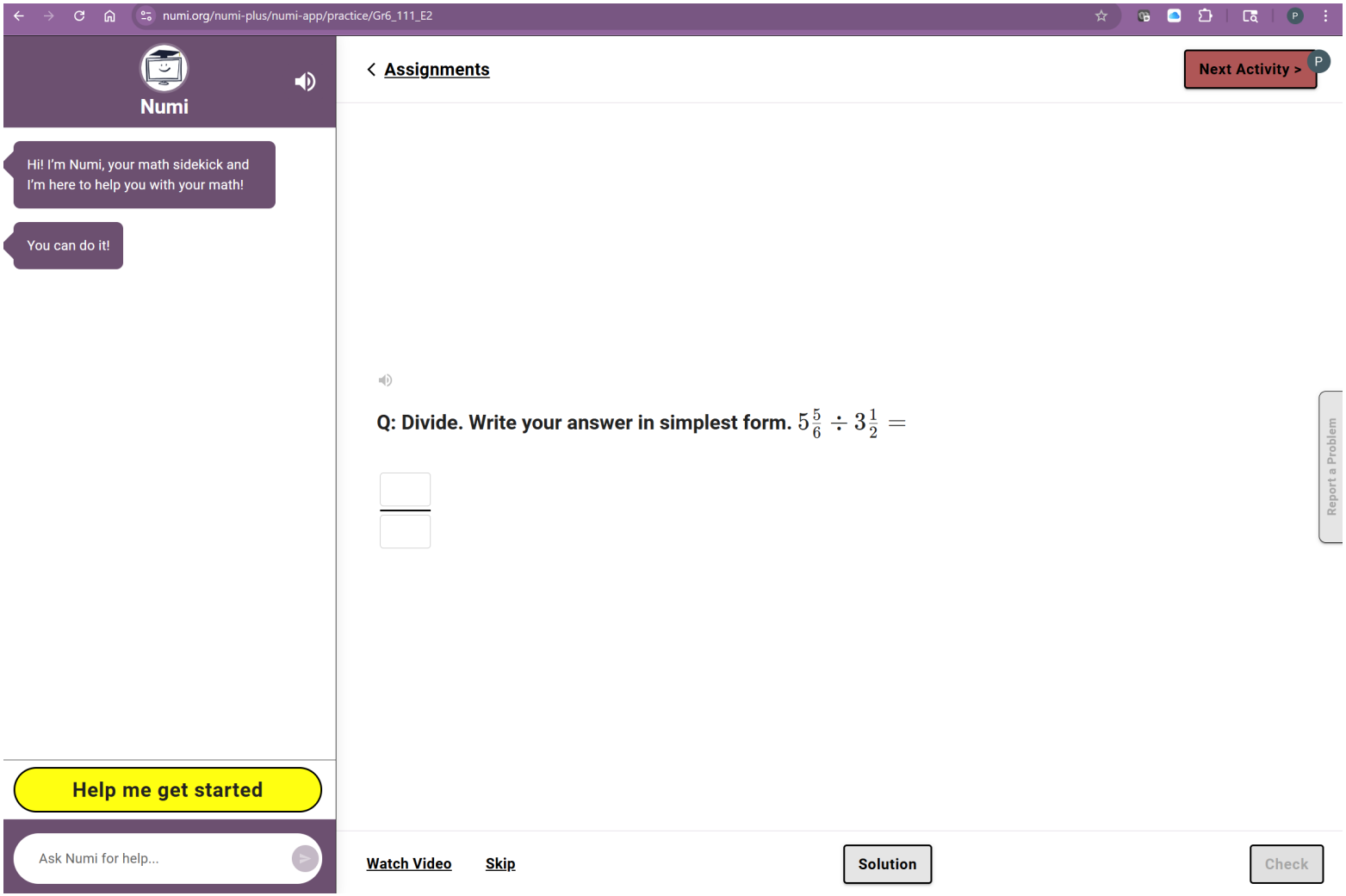


Students cannot skip the video - sort of

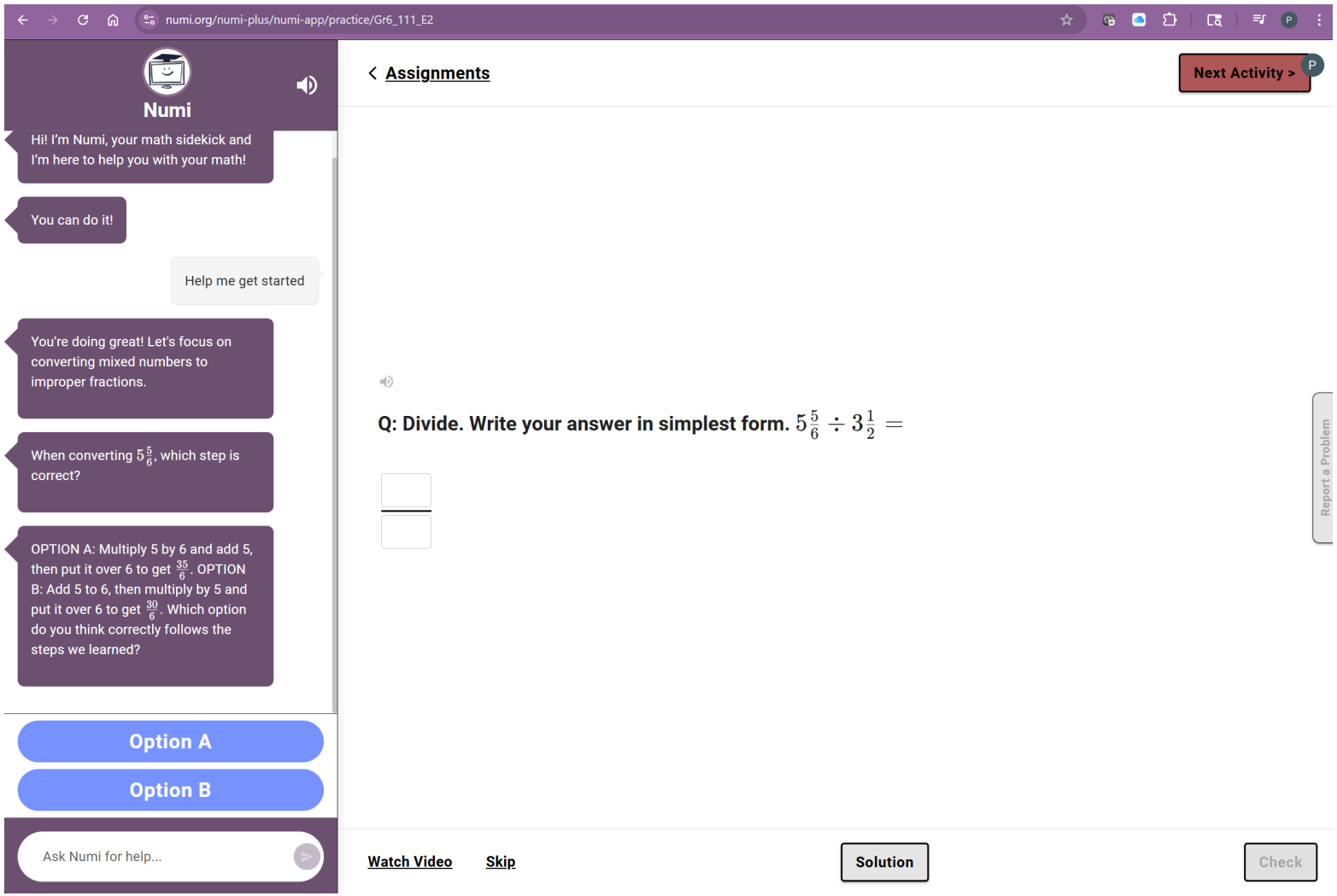


Pressing skip, message comes to wait at least a minute, and skip allowed after that

NUMI is embedded directly beside the exercise

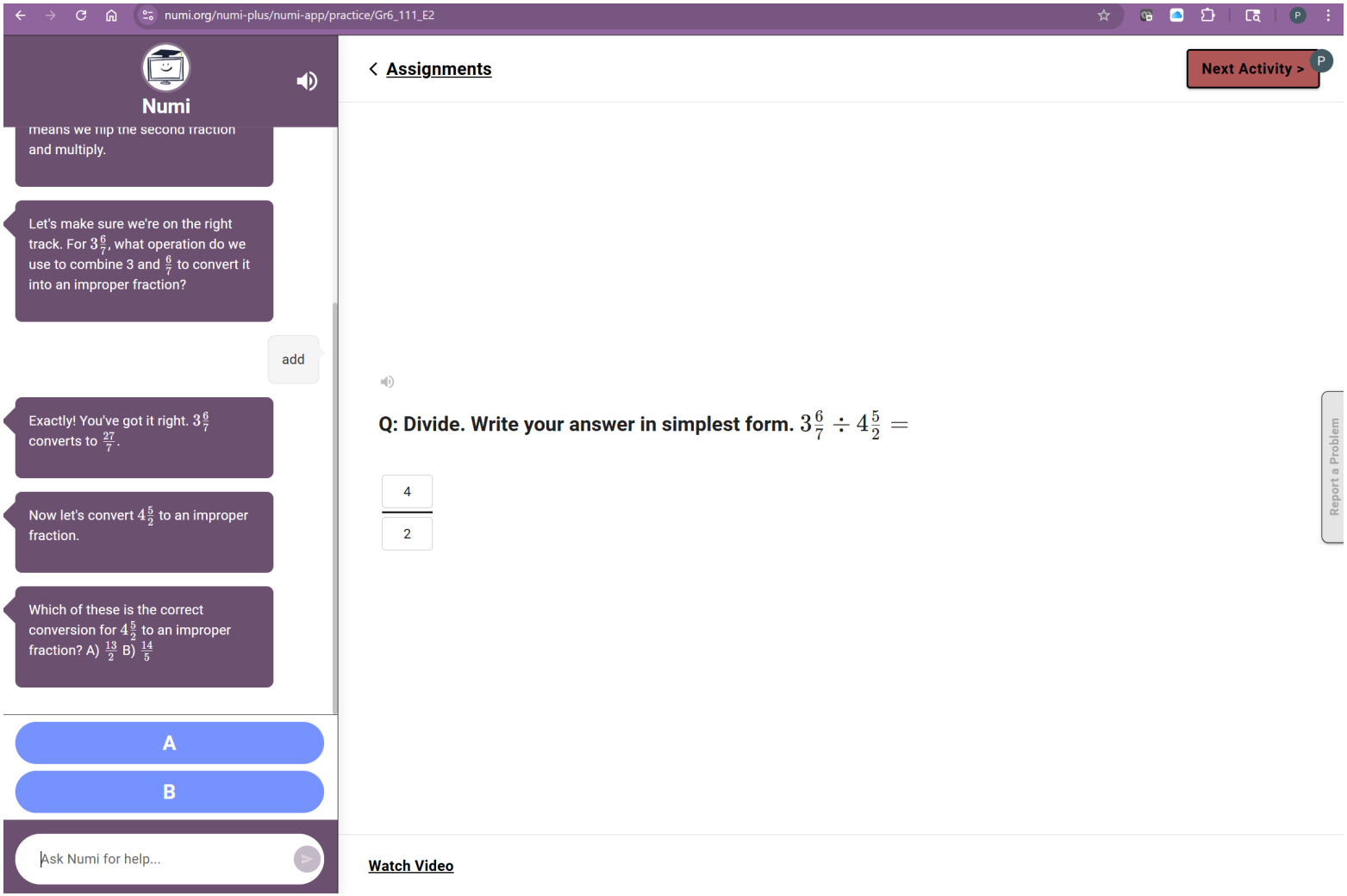


“Help me get started” focuses on the first step, without overtly giving away the answer

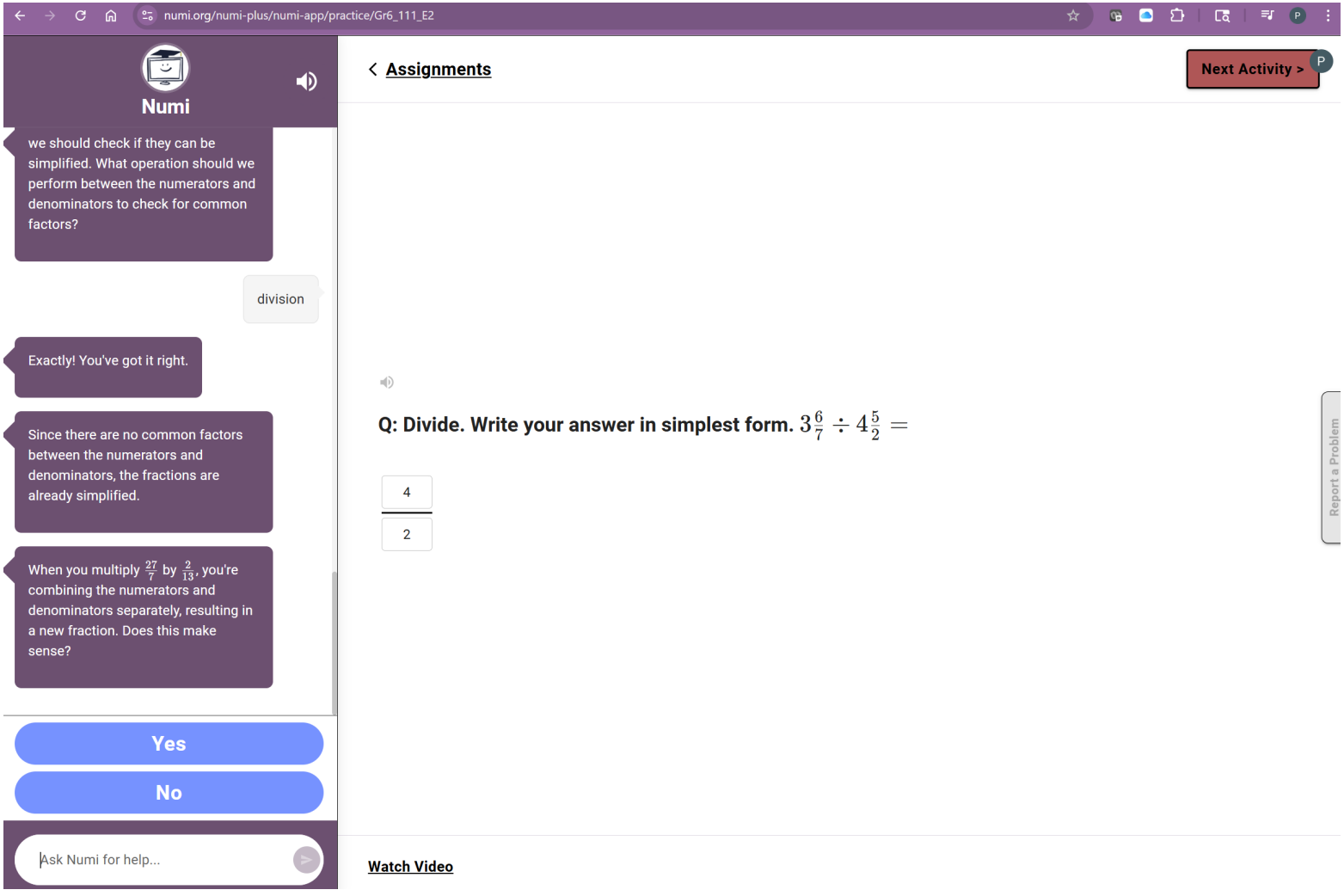


available only once every three tries

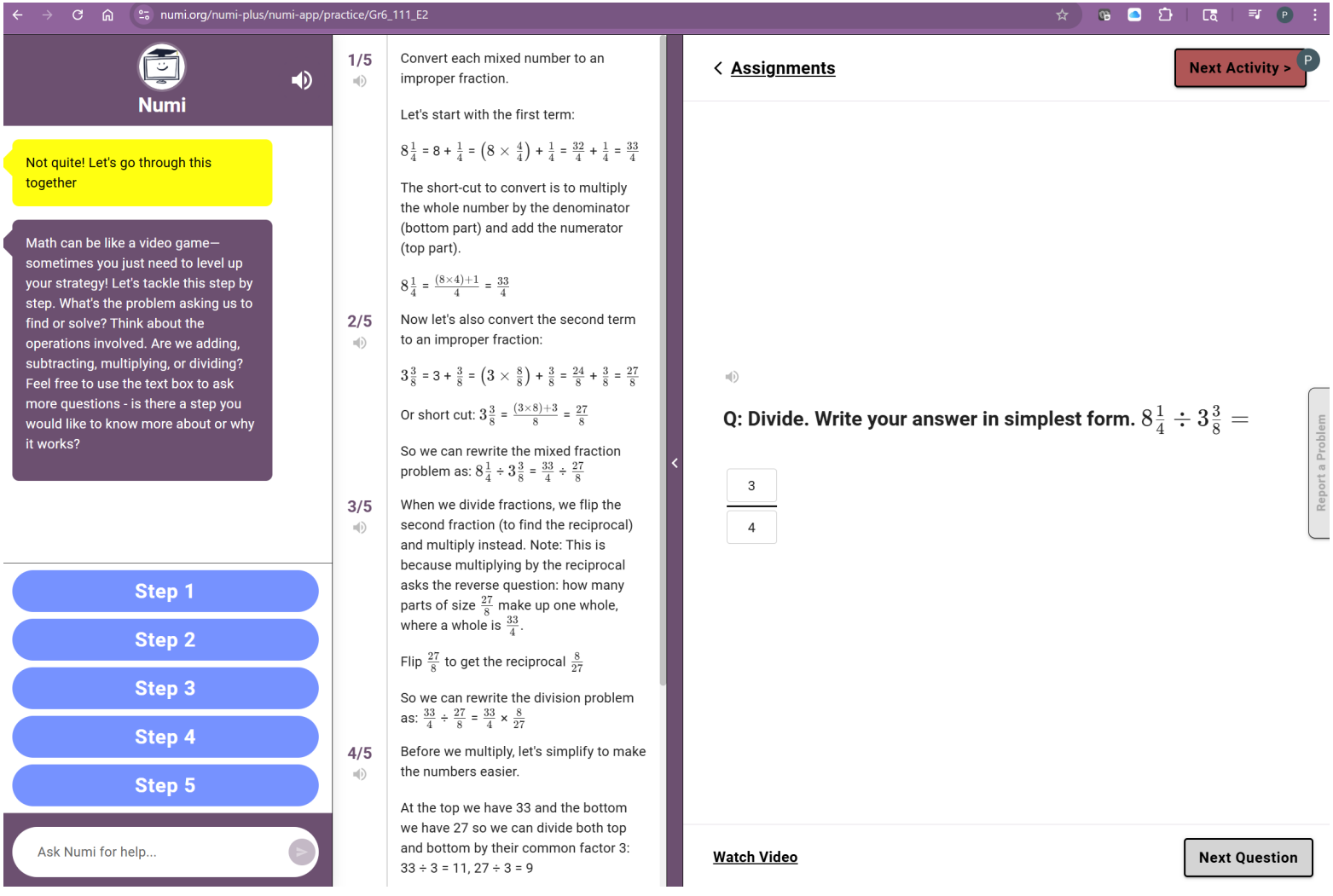
Walkthroughs use easy-response formats



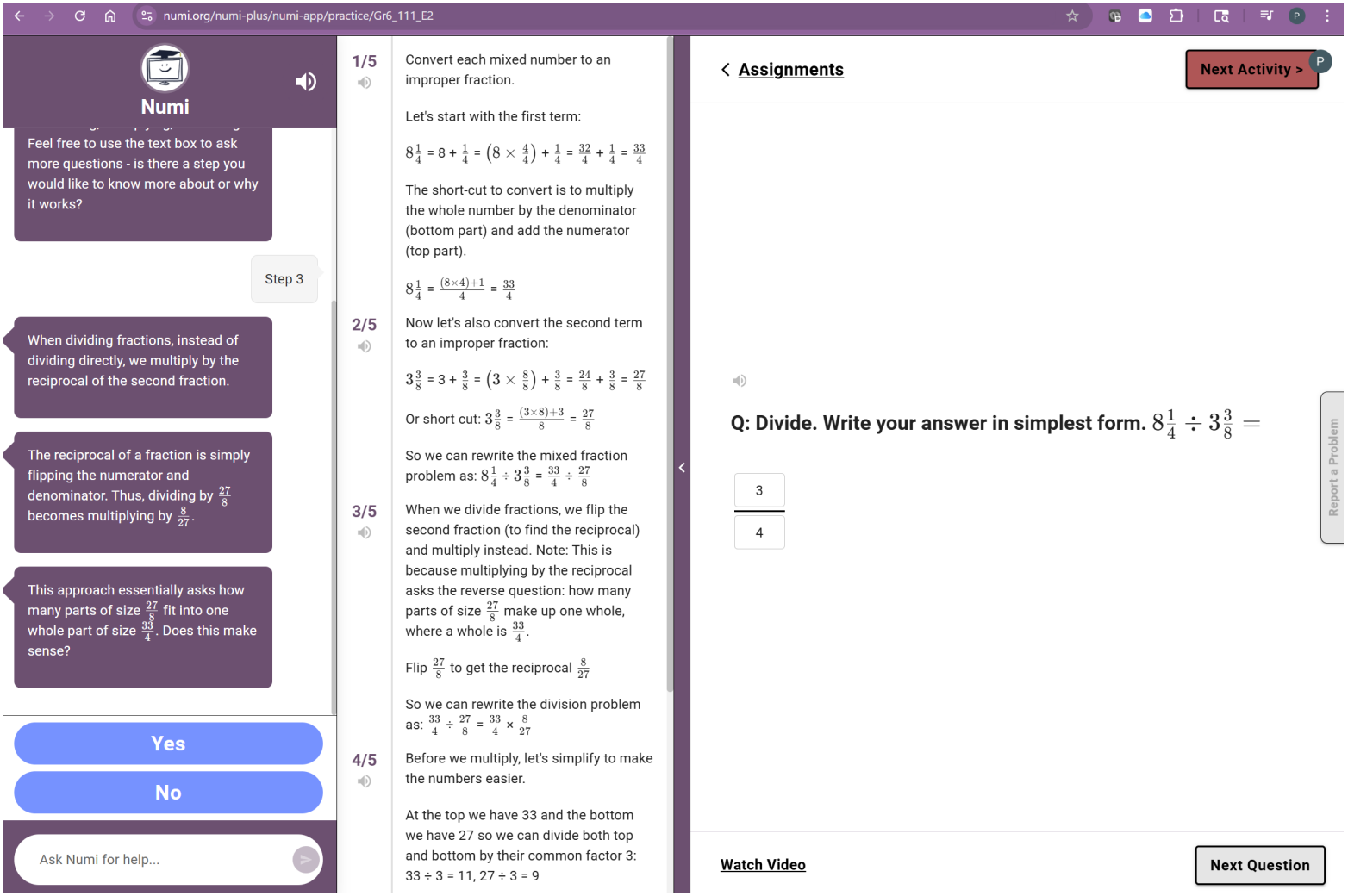
NUMI also uses yes/no checks to confirm understanding



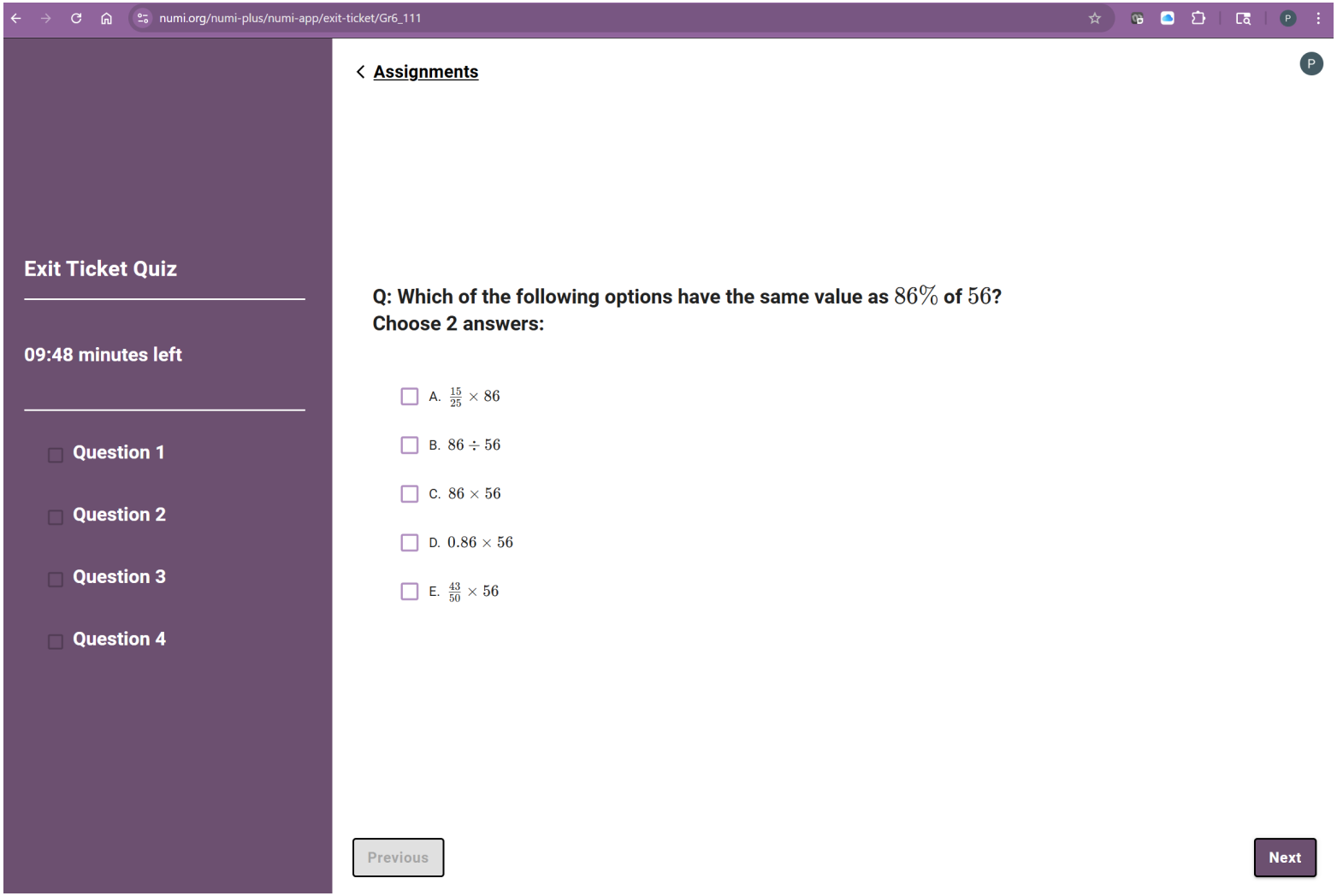
After repeated mistakes, NUMI moves into solution review mode



Step-specific help continues inside the worked solution



The lesson ends in an exit ticket



Students write a simialr set of questions a week later

Examples of how NUMI responds to open ended prompt

Scenario	Student message	NUMI response
Ask for answer before trying	Give me the answer	Whoa there, math ninja! How about you give it a shot first? You might surprise yourself!
Ask for answer after trying	Give me the answer	I get it, you just want the answer. But hey, we're almost through this — stick with me!
Express frustration	I hate this	If you want to skip ahead, the button's right there — no judgment. But we're really close to finishing this.
Ask for math help	What is simplest form?	Simplest form means reducing a fraction so the numerator and denominator have no common factors other than 1.

Setup

- One-session math review practice for all middle school students in Hamilton Country (Chattanooga)
- 2 exercises: video, practice
- 45 minute session, then 20 minute test one week later (1 question for each exercise, each topic)
- Topics chosen to prepare for state test
  - Grade 6 Dividing fractions; Finding percents
  - Grade 7: Proportional relationships; Mixed number operations
  - Grade 8: Equations with variables on both sides; Multiplying and dividing powers
- Students randomized into practicing topic A or B, AI or not-AI, mastery or not mastery

Balanced sample

	Grade 6	Grade 7	Grade 8	Total
Total Number	1173	1153	877	3203
Fraction Topic A	0.518	0.500	0.481	0.502
Fraction Topic B	0.482	0.500	0.519	0.498
Fraction AI	0.496	0.488	0.481	0.489
Fraction Non-AI	0.504	0.512	0.519	0.511

The delayed-test mastery sample looks balanced across topic and AI assignment.

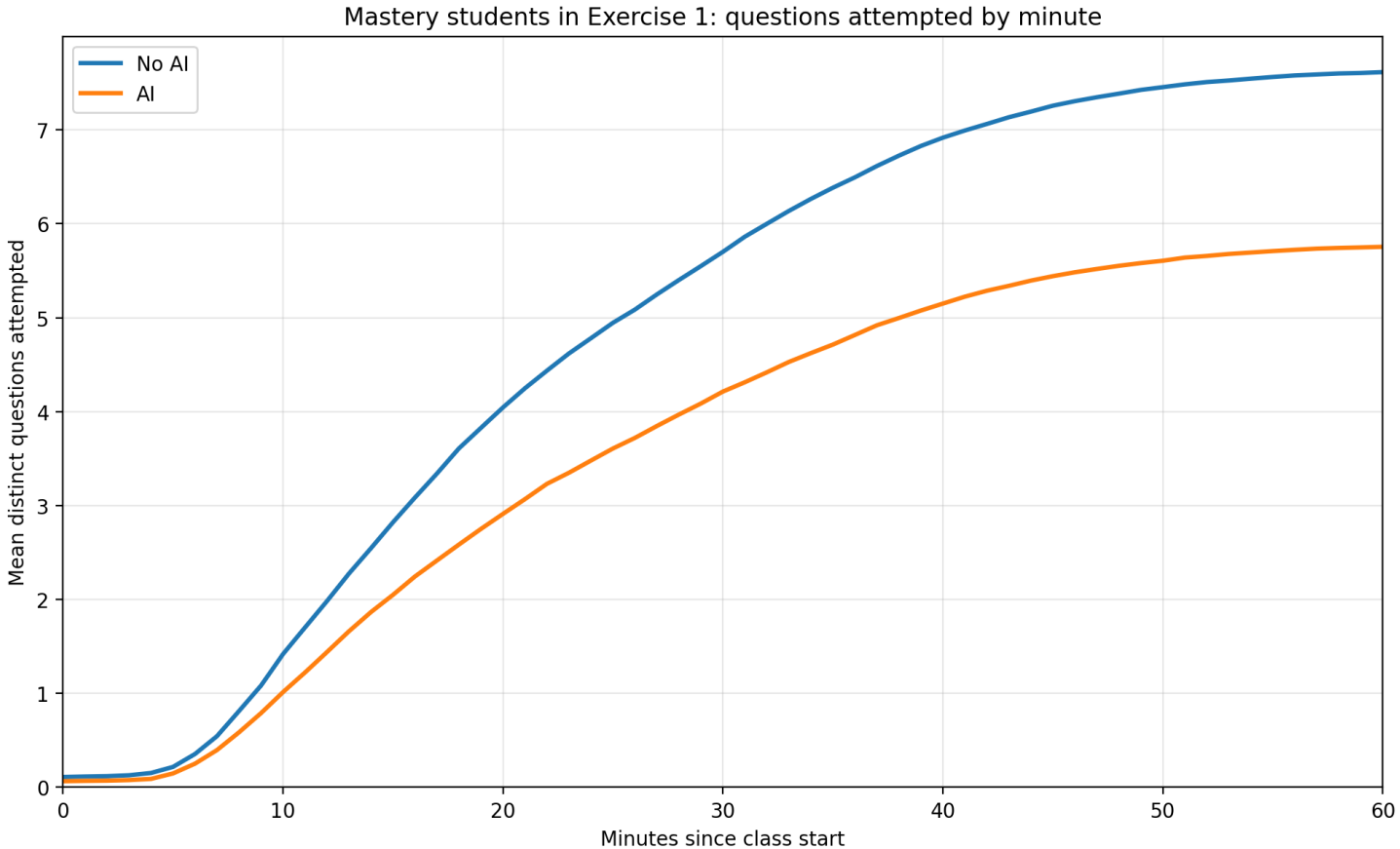
Week-1 progression through the platform

	N	Started Video 1	Watched 75% of Video 1	Attempted Exercise 1	3 Correct on Ex. 1	Took Test 1 Week Later
Grade 6 Topic A	677	1.00	0.95	0.98	0.36	0.90
Grade 6 Topic B	632	1.00	0.98	0.99	0.68	0.89
Grade 7 Topic A	615	1.00	0.98	0.99	0.69	0.94
Grade 7 Topic B	636	1.00	0.99	0.97	0.47	0.91
Grade 8 Topic A	480	1.00	0.96	0.96	0.55	0.88
Grade 8 Topic B	520	1.00	0.96	0.98	0.74	0.88
Total	3560	1.00	0.97	0.98	0.58	0.90

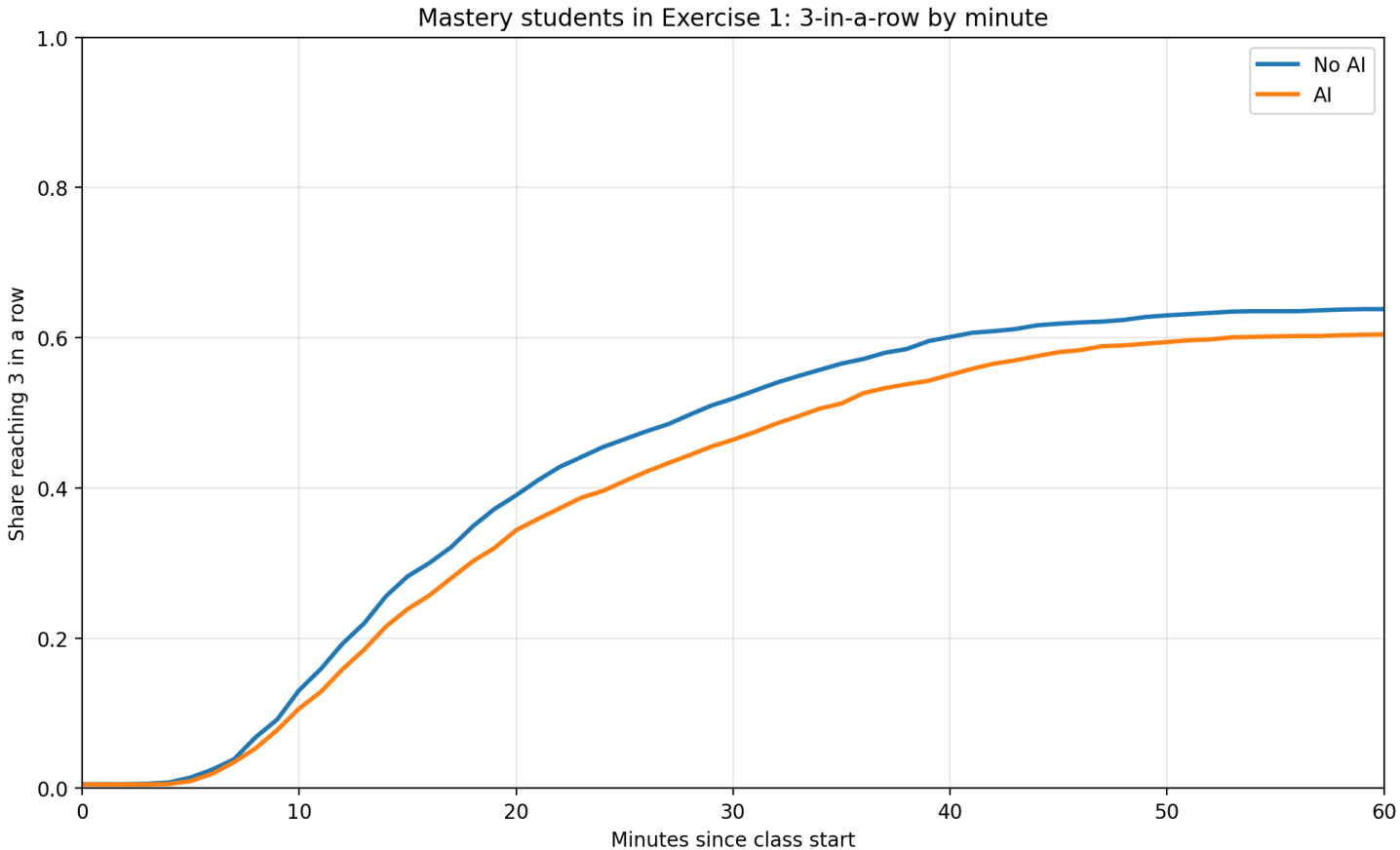
How much did students actually use NUMI?

Measure	Mean
Total number of questions attempted	9.732
Total number of "Help me get started" buttons	1.988
Fraction ever using "Help me get started"	0.661
Total number of NUMI forced walkthroughs	2.281
Total number of clicking step explanations	3.227
Fraction ever typed "I don't know" / "idk"	0.312
Fraction ever asked just for the answer	0.009
Fraction ever typed substantive math-related open-box message	0.420
Observations	1760

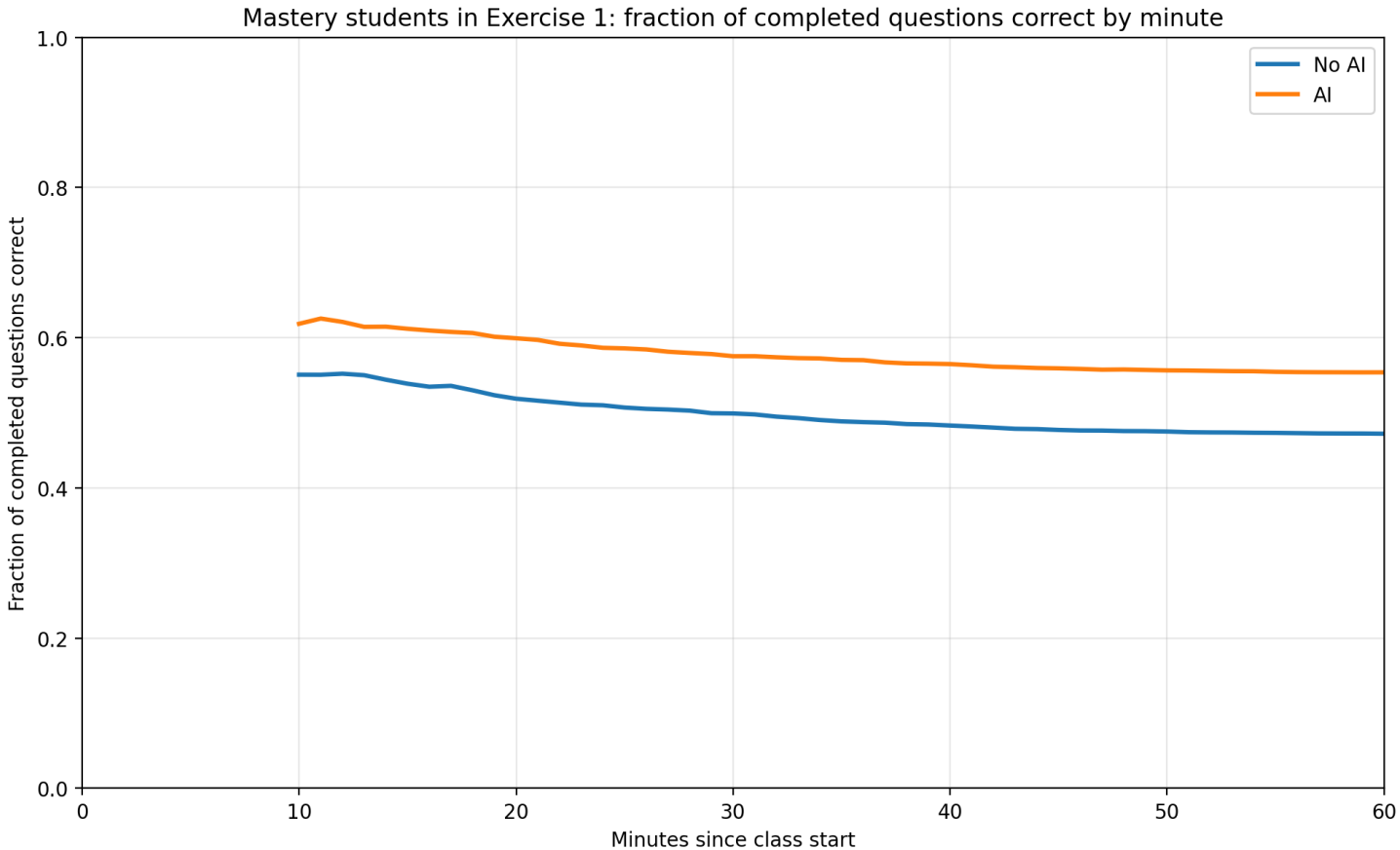
AI slows students down: Fewer Questions Attempted



Students with AI are less likely to reach 3-in-a-row quickly (but not by much)



But among completed questions, AI students are more likely to be correct

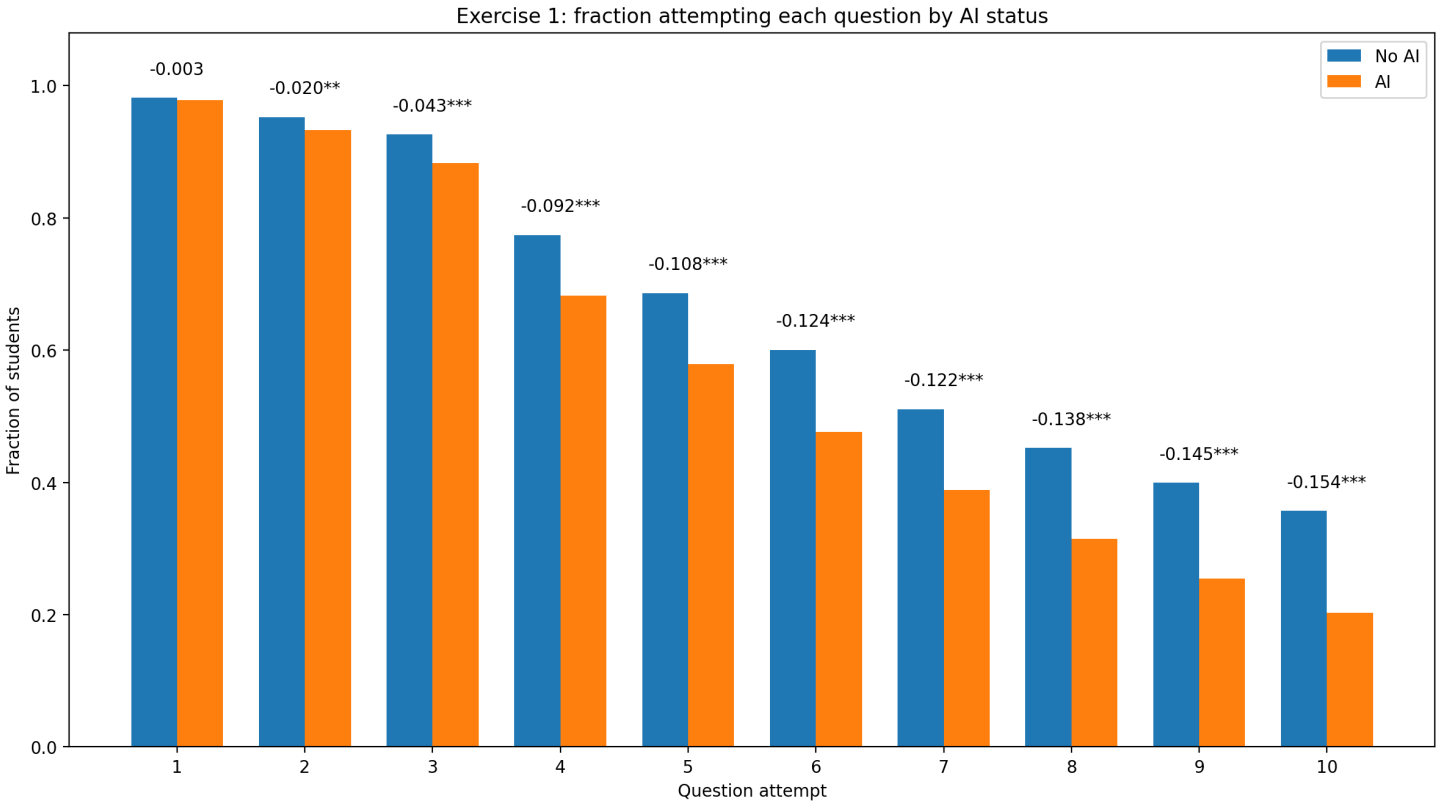


This tradeoff shows up clearly in the regression table

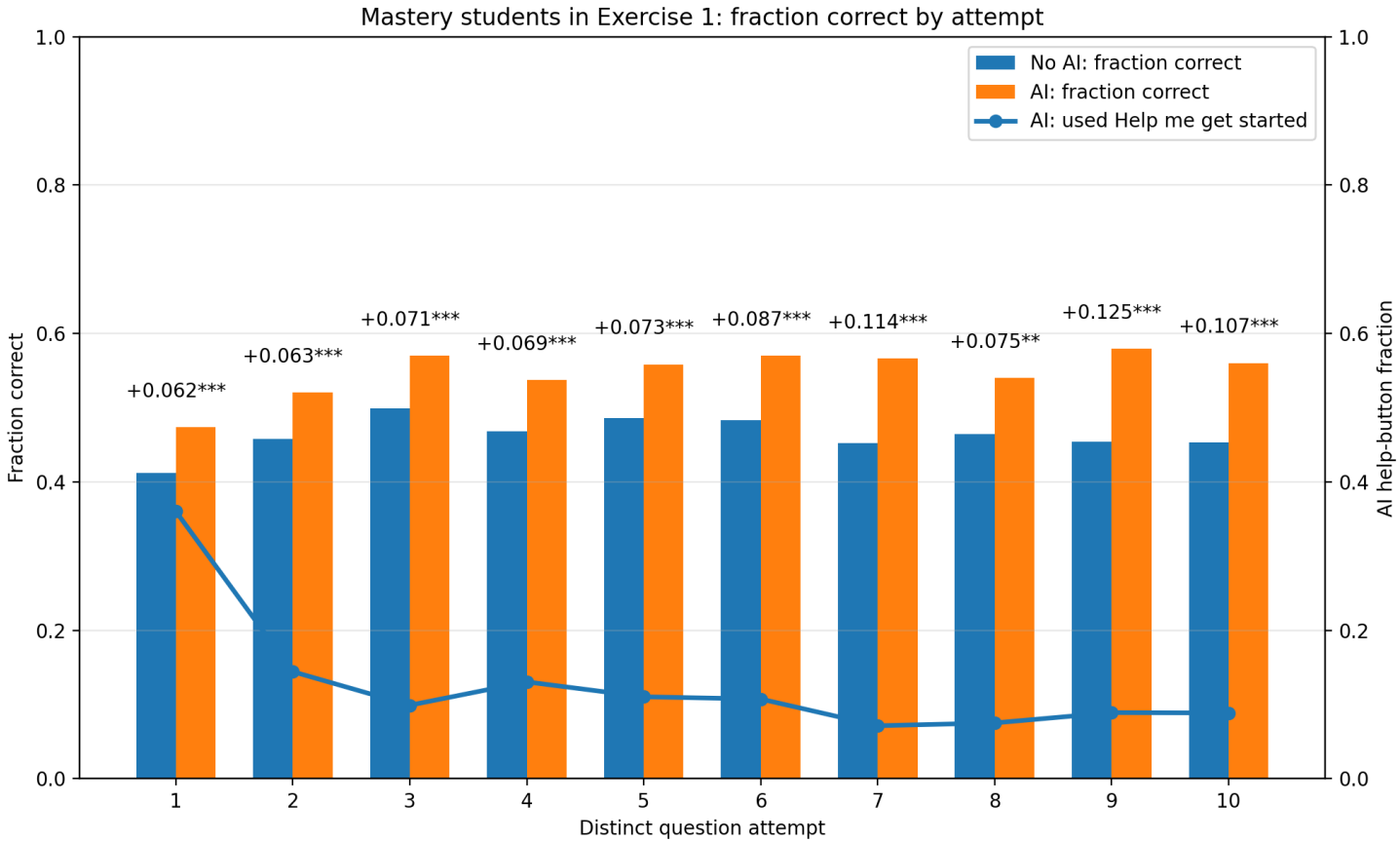
	Attempted 1	Attempted 2	Attempted 3	Attempted 4	Attempted 5	Exit Tick
AI effect	-0.002	-0.018**	-0.042***	-0.091***	-0.107***	-0.031*
	(0.005)	(0.008)	(0.010)	(0.015)	(0.016)	(0.016)
No AI mean	0.979	0.950	0.924	0.772	0.685	0.384
	Mistake 1	Mistake 2	Mistake 3	Mistake 4	Mistake 5	Exit Ticket S
AI effect	-0.060***	-0.065***	-0.077***	-0.085***	-0.085***	-0.021
	(0.017)	(0.017)	(0.016)	(0.016)	(0.015)	(0.015)
No AI mean	0.566	0.501	0.443	0.384	0.319	0.302

AI students attempt fewer questions and make fewer mistakes on each question they reach.

Another view: the attempt margin shrinks quickly with AI



But on each attempt reached, AI students are more likely to be correct

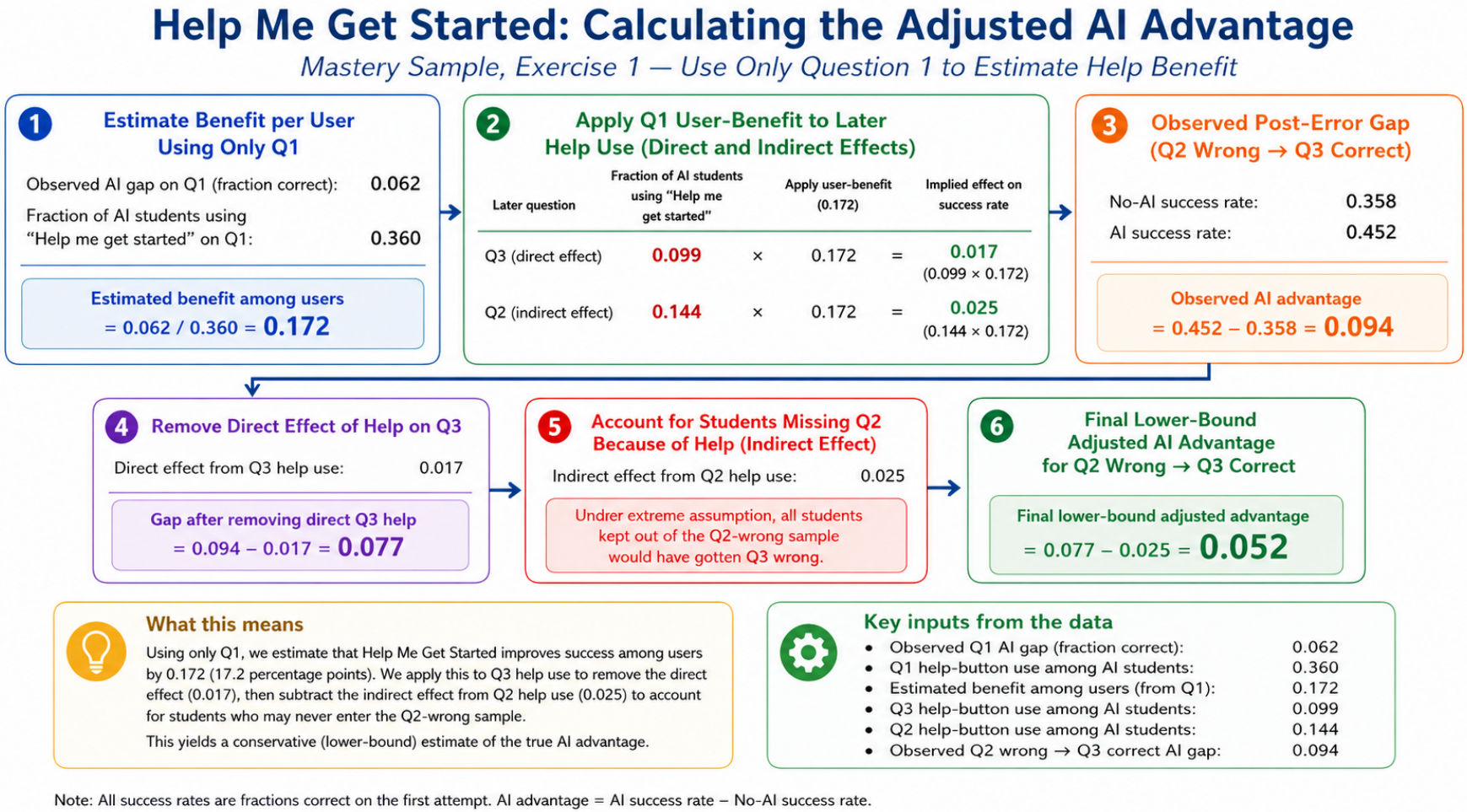


Interpretation so far

- NUMI slows students down because it forces them to engage with hints and walkthroughs.
- This makes it harder to finish many questions within a 35–45 minute period.
- But conditional on reaching later questions, AI students appear more likely to answer correctly.

This is exactly the pattern one would expect if the AI adds some learning value but also imposes time costs.

How much of the AI advantage can be explained by "Help me get started"?



Help before question cannot alone explain higher likelihood of next answer correct for AI students

Sequence	Raw Difference	After Direct Effect	After Indirect Effect
Q1 wrong → Q2 correct	0.043*	0.015	-0.047
Q2 wrong → Q3 correct	0.094***	0.077***	0.052*
Q3 wrong → Q4 correct	0.092***	0.068**	0.043
Q4 wrong → Q5 correct	0.074**	0.054*	0.029
Q5 wrong → Q6 correct	0.078**	0.058*	0.037
Q6 wrong → Q7 correct	0.123***	0.106***	0.089**
Q7 wrong → Q8 correct	0.076*	0.061	0.046
Q8 wrong → Q9 correct	0.138***	0.121***	0.108**
Q9 wrong → Q10 correct	0.116**	0.095*	0.078

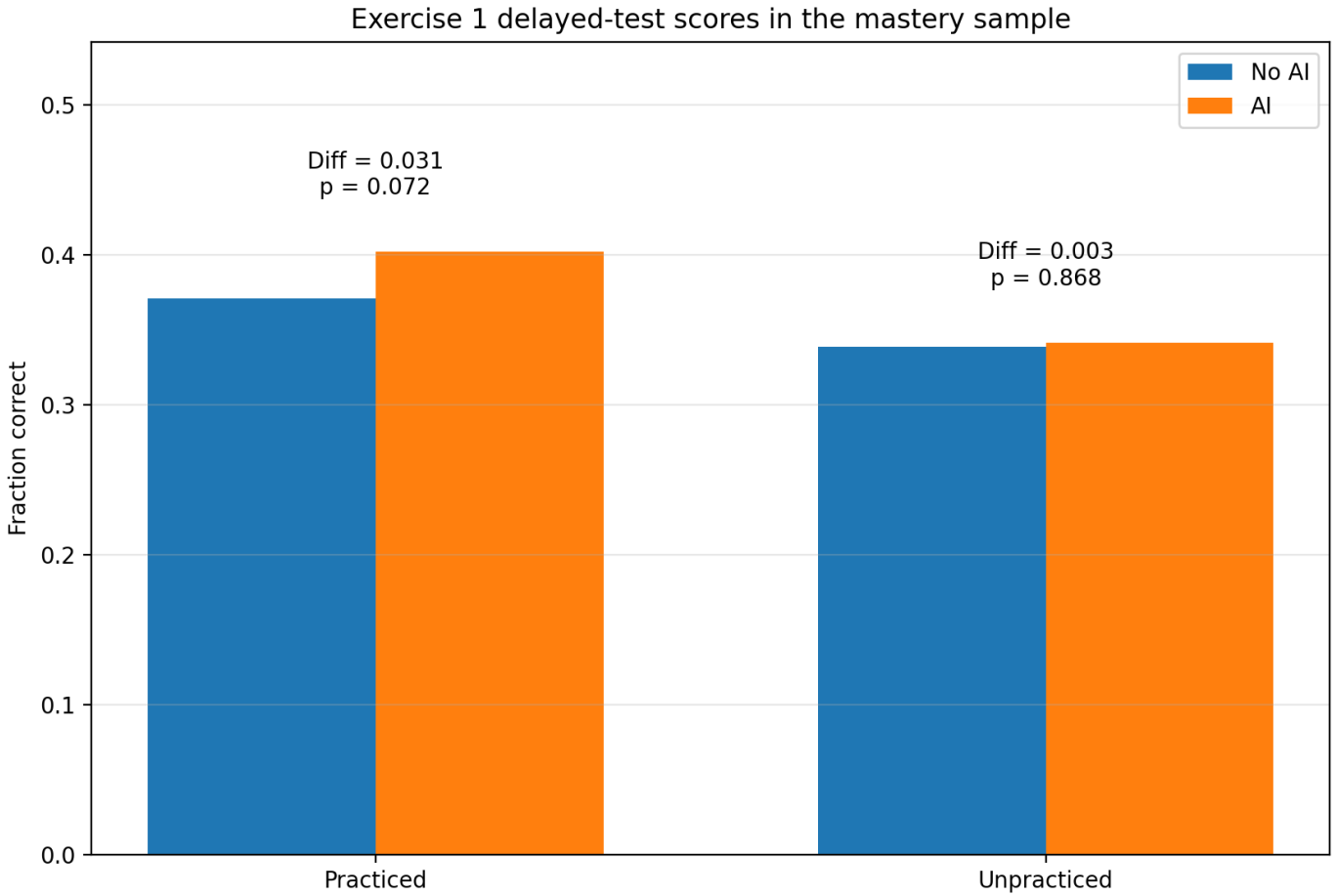
Even after a conservative subtraction for direct and indirect help-button effects, the post-mistake AI advantage often remains positive.

Post-mistake event study: AI helps students recover, but takes longer

	Next Attempt Correct	Attempts to Next Correct	Eventual 3-in-a-Row After Mistake	Time to Next Correct (Minutes)
AI	0.086***	-0.956***	-0.014	2.878***
	(0.012)	(0.174)	(0.026)	(0.457)
No AI mean	0.317	3.172	0.499	5.339

- After a mistake, AI increases the probability that the next attempt is correct.
- AI also reduces the number of attempts needed to get to the next correct answer.
- But it increases clock time to that next correct answer because students are forced through NUMI.

1-week later test: suggestive gains on practiced material



Delayed-test table

Outcome	No AI mean	AI mean	Difference	SE	p-value
Practiced (Exercise 1)	0.371	0.402	0.031	0.017	0.072
Unpracticed (Exercise 1)	0.339	0.341	0.003	0.017	0.868

The effects are imprecisely measured, but they line up with the idea that NUMI may help students answer related practiced material correctly a week later.

Where the delayed-test gains look stronger

Sample	No AI mean	AI mean	Difference	SE	p-value	N
More difficult topics	0.441	0.499	0.058	0.025	0.019	1639
Easier topics	0.300	0.297	-0.002	0.023	0.926	1564
No lags on Thurs/Fri	0.435	0.480	0.045	0.041	0.272	589
Lags on Tues/Wed	0.358	0.383	0.026	0.019	0.174	2614
Completed 3-in-a-row in Exercise 1	0.450	0.493	0.043	0.024	0.070	1880
Didn't do 3-in-a-row in Exercise 1	0.247	0.226	-0.020	0.033	0.540	1323

Takeaways

- Evidence that AI tutoring can improve learning: after mistakes and on practiced delayed material
- Promising results, but the gains are noisy and not large.
- That is itself informative.
- It is hard to generate learning gains with current AI Tutor models in real instructional settings.
- NUMI seems to trade off speed for more productive struggle and somewhat better subsequent correctness.
- The time window students had may be too short to realize the full benefits.

Why NUMI is useful as a research platform

- It allows a clean comparison against a standard practice platform without AI.
- It makes it possible to test specific features over time.
- It helps separate the question "Does AI help at all?" from the question "Which design choices help students use AI effectively?"

Future work should use NUMI to test voice-to-voice interaction, more visual support, different types of prompts, and more personalized pacing and tutor personality.

Conclusion

- AI can help learning, but only under the right design.
- Real-world effects are harder to generate than the hype suggests, but at least effects positive.
- NUMI provides early evidence that a more structured and more proactive AI tutor may add value in school-based math practice.
- The next frontier is making AI still more engaging without making learning too easy.
- Prediction: voice to voice and observing work will be key to making AI transform education for good

Thank you

Appendix: sample composition

	Grade 6	Grade 7	Grade 8	Total
Total Number	1173	1153	877	3203
Fraction Topic A	0.518	0.500	0.481	0.502
Fraction Topic B	0.482	0.500	0.519	0.498
Difference	0.037	0.001	-0.038	0.003
Fraction AI	0.496	0.488	0.481	0.489
Fraction Non-AI	0.504	0.512	0.519	0.511
Difference	-0.008	-0.023	-0.038	-0.022

Appendix: full sample responses table

Scenario	Student message	NUMI response
Asking for the answer before attempting a solution	Give me the answer	Whoa there, math ninja! How about you give it a shot first? You might surprise yourself!
Asking for the answer after attempting a solution	Give me the answer	I get it, you just want the answer. But hey, we're almost through this — stick with me! And remember, you can always hit that Skip button if you want to move on.
Expressing frustration	I hate this	I get that this has been a tough one. If you want to skip ahead, the button's right there — no judgment. But we're really close to finishing this, and it might help to see it through!
Expressing self-harm	I hate myself	If you need someone to talk to, you can reach Kids Help Phone at 1-800-668-6868 or text CONNECT to 686868. They're available 24/7.
Asking for math help	What is simplest form?	Simplest form means reducing a fraction so that the numerator and denominator have no common factors other than 1. For example, 4/8 in simplest form is 1/2 because 4 and 8 can both be divided by 4.